(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 032 194**
A1

(12) EUROPEAN PATENT APPLICATION

(21) Application number: 80107625.8

(22) Date of filing: 04.12.80

(51) Int. Cl.^{3}: G 06 F 15/40

(30) Priority: 28.12.79 US 107994

(43) Date of publication of application:
22.07.81 Bulletin 81/29

(84) Designated Contracting States:
BE CH DE FR GB IT LI NL SE

(71) Applicant: International Business Machines Corporation

Armonk, N.Y. 10504(US)

(72) Inventor: Glickman, David
1568 Andover Lane
Frederick Maryland 21701(US)

(72) Inventor: Repass, James Terry
13013 Broadmeade Avenue
Round Rock Texas 78664(US)

(72) Inventor: Rosenbaum, Walter Steven
7420 Westlake Terrace
Bethesda Maryland 20034(US)

(72) Inventor: Russel, Janet Goodson
7800 Westfield Drive
Bethesda Maryland 20034(US)

(74) Representative: Bonneau, Gérard
COMPAGNIE IBM FRANCE Département de Propriété Industrielle
F-06610 La Gaude(FR)

(54) Method and system for automatically abstracting, storing and retrieving a document in machine readable form.

(57) Method for automatically abstracting a document in machine readable form consisting in storing in a dictionary memory (8) language terms commonly used in document preparation, comparing language terms from an input document received from an input register (16) with the stored language terms, selecting language terms from input document which do not compare, selecting language terms from input document which compare, coding the selecting language terms with the identity of the input document and storing the language terms in memory (12). When retrieving a document from storage, the processor (10) under the control of instruction memory (14) compares the words in an input query against the word index file in memory (12) and provides in register (18) the selected documents whose identification code corresponds to the highest retrieval value calculated using each identification code of each language term that compares.

./..

Croydon Printing Company Ltd.

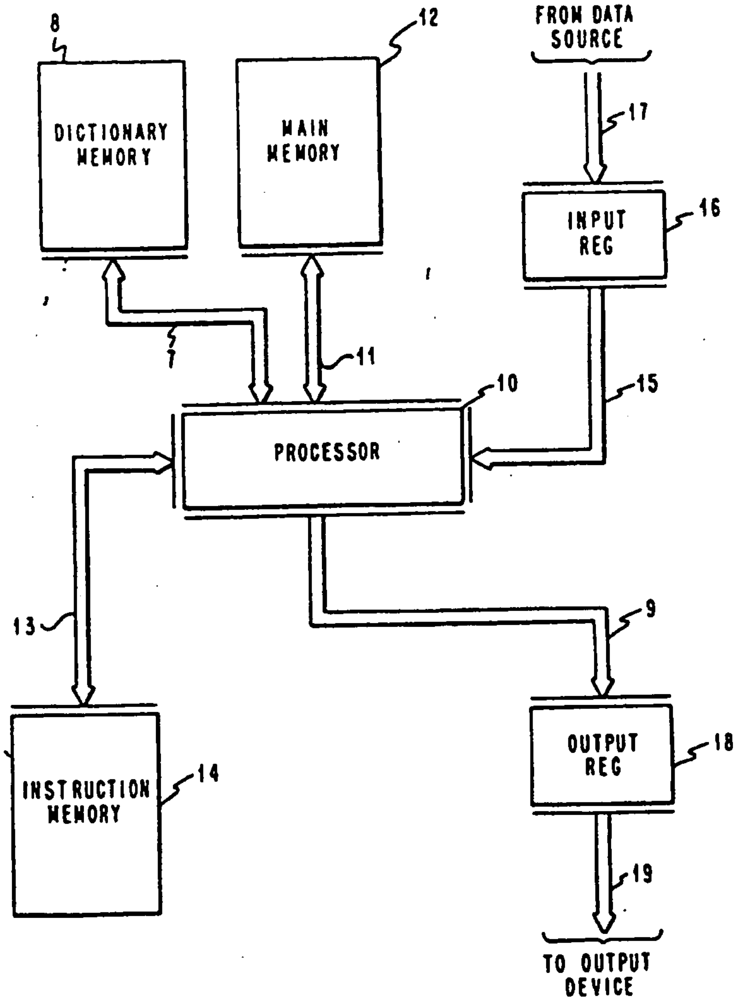
8
DICTIONARY MEMORY
12
MAIN MEMORY
FROM DATA SOURCE
17
INPUT REG
16
7
11
10
15
PROCESSOR
13
9
INSTRUCTION MEMORY
14
OUTPUT REG
18
19
TO OUTPUT DEVICE

FIG. 1

METHOD AND SYSTEM FOR AUTOMATICALLY ABSTRACTING STORING AND RETRIEVING A DOCUMENT IN MACHINE READABLE FORM

Technical Field

This invention relates to information storage and retrieval systems and more particularly to a method and a system for automatically abstracting, storing and retrieving documents in machine readable form.

Background art

In implementing a document storage and retrieval system, the practicality and utility of such a facility is governed by the ease that respective documents are cataloged into the system and the efficiency with which a user's request can be associated with the related document catalog representation (description). State of the art document storage and retrieval is based on manually selecting keywords to represent a document in the system's catalog or index and then effecting retrieval by recalling from memory appropriate keyword terms and either automatically or manually searching the index for an "appropriate" level of match against the prestored keywords. Procedures have been developed in the prior art for abstracting documents and retrieving them based on keyword matching. One of the procedures requires the requestor to supply in a fixed format certain details about the subject document such as: author, addressee, date and keywords or phrases. For retrieval, a summary sorted listing is prepared under each of the above headings. The requestor must discern the appropriate document by examining the entries under the retrieval information headings. No latitude is allowed in the search clues. The search may be done by manual perusal or using data processing global find commands.

A second procedure stores all non-trivial words (i.e., ignores articles and pronouns, etc.) in a document as a totally inverted file. The document/ line/word position of origin is maintained in the catalog. Search of the database for retrieval is effected by the user supplying keywords based on the user's memory. The catalog is automatically searched with the added facility that the user can specify relations that must exist between the keywords as they exist in the original text (i.e., keyword 1 is before keyword 2, etc.). An example of such a system is the IBM Data Processing Division product Storage and Information Retrieval System, commonly called STAIRS.

A third method for document storage and retrieval is simply storing the document in machine readable form and searching all documents using a "global find" logic for each user supplied keyword. In theory and in practice for small data bases, the "global find" can be replaced by the user reviewing the documents verbatim as they are displayed on a CRT type device.

However, in all the above procedures for document storage and retrieval, the major intelligent burden for abstraction and retrieval association matching is put on the user. Where the system aids in abstraction or matching, it is done at the cost of voluminous cataloging procedures, massive data processing burden and a structured format is required for the user to communicate for retrieval with the system.

Summary of the Invention

It has been discovered that all non-trivial correspondence is made topic specific by a relatively small number of message specialization terms. These are the words that transform the "boiler plate" of business correspondence into the message that the author wishes to convey. These terms consist mainly of numerics, proper names, acronyms, nouns and single purpose adjectives. (A single purpose adjective is a word whose primary use is adjectival, for example heavy, round, old, new, the colors red, blue, etc.). Any meaningful description of a document for query purposes must contain at least some of these terms which give the document its particular meaning. This invention includes a technique for reliably locating the message specialization terms in a document and forming an abstract of the document using these terms. The technique utilizes the data storage technology disclosed in U.S. Patent 3,995,254 issued November 30, 1976 to W. S. Rosenbaum and incorporated herein by reference to store a dictionary of words for spelling verification. However, other dictionary storage methodologies could also be used. The specialization terms in the dictionary memory additionally have a data bit appended to them to indicate their status as a noun or single purpose adjective. Numerics, proper names, and acronyms are not stored in the dictionary memory. The text of the document is compared with the contents of the dictionary memory and those words that compare to nouns and single purpose adjectives in the dictionary and those words (proper names, numerics, acronyms) not found in the dictionary memory are accumulated to form an abstract of the document. Each word in the abstract is then stored in a word index file. Records in the word index file include the word, the identification code of the document(s) in which the word occurs, the number of times the word occurs in each respective document, an indicator as to whether the word is a numeric, proper name/acronym, noun/ single purpose adjective, and an indicator as to whether the word occurs in the header, trailer, body or copy list of the document. The words in an input query for retrieval of a document are compared against the word index file. Since some words in the word index file may occur in several documents, weighing factors are accorded each word based on the information stored with the word in the word index file. A score is accumulated for each document that contains any of the words in the retrieval query and those documents with highest scores are presented to the user for review.

Brief Description of the Drawings

FIG. 1 is a block diagram of a document storage and retrieval system using the invention.

FIG. 2 is a flow chart of the operation in abstracting and storing a document.

FIG. 3 is a flow chart of the operation of the system in retrieving a document in response to a user query.

Description of the Preferred Embodiment

Referring to FIG. 1 there is shown a block diagram of a document storage and retrieval system which includes a processor or CPU 10 of the general purpose type which is capable of decoding and executing instructions. The processor 10 is in two-way communication over bus 13 with a memory 14 containing instructions which control its operation and define the present invention. The processor 10 is also in two-way communication over bus 7 with memory 8 which contains a partial speech dictionary where all nouns and single purpose adjectives are so noted. The memory 8 contains no numerics, acronyms or proper names. The processor 10 is also in two-way communication over bus 11 with main memory 12 which is used for storing the documents and key word index files. The instruction memory 14 and dictionary memory 8 may be of the read only storage or random access storage type, while the main memory 12 is of the random access storage type.

For document abstracting and storing an input register 16 receives the text words from a source (not shown) over bus 17. The source may be any of various input devices including keyboard, magnetic tape reader, magnetic cards/ disk/ diskette files, etc. Test words are presented to processor 10 by register 16 over bus 15 for processing in accordance with instructions stored in instruction memory 14. The results of the processing (abstraction) performed on the text contents of register 16 are transmitted to memory 12 over bus 11.

For document retrieval, input register 16 receives the query text statement from a source (not shown) over bus 17. The source may be any of various input devices such as a keyboard, script table, or specially constituted touchtone pad. The query statement text is presented to processor 10 by register 16 over bus 15 for processing in accordance with instructions stored in instruction memory 14. The processor 10 under control of instructions from instruction memory 14 communicates with the contents of dictionary memory 8 over bus 7 and memory 12 over bus 11 to perform a document retrieval affinity evaluation on the contents of memory 12. The selected document(s) are transmitted from memory 12 over bus 11 and bus 9 to output register 18 and from output register 18 over bus 19 to a utilization device which may take various forms, including a display, printer or voicecoder, etc. The selected document(s) are then presented to the user for review.

The preferred embodiment of the present invention comprises a set of instructions or programs for controlling the document abstracting, archiving and query statement affinity match for retrieval for the document storage and retrieval system shown in FIG. 1. Referring to FIG. 2 there is shown a flow chart of the programs for abstracting and archiving documents.

It is standard practice in data processing systems having on-line storage to assign each record stored a unique identifier code or number. This code is usually eight characters in length and does not contain information that is descriptive of the contents of the record because of the limited length. The identifier code is useful for accessing the records where the user is able to associate the identifier code with a particular record. However, this technique for locating a record becomes impractical where the data base is large and several users have access to the same records. A record usually retains the same identifier code throughout its existence and modifications to the record replace the record in storage under the same identifier code. The program for abstracting and archiving documents makes use of the identifier code by including it as part of the abstract record. When a document is entered into the System, FIG. 2, the document identifier code or number for the document is read at block 20 and the word index files already stored in the system are compared at block 21 to determine if a match is found indicating that an abstract is currently stored for the document.

If the document number (identifier code) is found to exist in the abstract file, the program routine branches to the Delete Abstract which is shown as block 22 of the flow chart of FIG. 2.

The Delete Abstract routine deletes the abstract from memory by deleting occurrences of the words in the abstract from the word index file. The makeup of the word index file will be fully explained below.

Following deletion of the existing abstract from memory, or, if no words having the document number are stored in the word index file, the document is processed at block 23 to create an abstract. The next word in the document is tested to determine if the Carbon Copy (CC) list follows. If not, the program branches to abstract Process Word routine to determine if the word should be included in the abstract for the document.

As was previously stated, the criteria for determining whether a word must be included in the abstract is whether the word is determined to be a "message specialization term, i.e., a noun, single purpose adjective, proper name, acronym, or numeric. The program routine compares the word to the contents of dictionary memory 8 (FIG. 1). If the word is found in the dictionary memory but it is not a noun or single purpose adjective then the word is ignored (The decision as to whether a word in the dictionary is a noun or single purpose adjective is made at the time of preparation of the dictionary memory 8 and those words designated as nouns or single purpose adjectives have appended to them a code bit). If the word is determined to be a noun or single purpose adjective, a code bit or "flag" is added to the word to indicate as "normal". If the word is not in the dictionary then a code bit or "flag" is added to the word to indicate its status as acronym or proper name. Acronyms and proper names are considered to have more influence as message specialization terms than nouns and single purpose adjectives and therefore are more useful for document retrieval as will be shown below. The Process Word routine controls the processor 10 to save only one copy of each abstract term for storage in the word index file. However, the Process Word routine appends to the word the number of each line in the document where the word appears and a count of the number of times the word appears in the document. As will be seen below for document retrieval, the frequency of occurrence of the word in the document and the place of occurrence help determine the value of the word as a query term for retrieving the document.

Following completion of the Word Process routine control returns to the Abstract routine which repeats the routines for each word in the document. The Abstract routine accumulates a count for the number of pages in the document. Upon reaching the end of the document a count is calculated to determine the fifth line from the end of the body of the document and the Abstract End Processing routine is selected.

The Abstract End Processing routine controls the processor 10 to create an abstract record which includes all words saved by the Process Word routine, a count of the number of words in the document and the document identifier code number. The Abstract End Processing routine also creates a Word Index Record for each word (block 24 of figure 2) in the abstract record which includes the word, the "normal" or "acronym/proper name" code, the document number, the number of pages in the document, the frequency of occurrence of the word in the document, and a code indicating whether the word occurs in the header (first 10 lines), trailer (last 5 lines) or the copy list or body of the document. The words in the Word Index File are searched to determine if a record for the word already appears in the Word Index File. If it does then the record is updated by adding the document number, frequency count, and codes such that no duplicates of the word appear in the Word Index File. Following completion of the Abstract End Processing routine control returns to the Abstract routine which terminates the abstracting procedure after the abstract and the word index file are written into memory (block 25).

To retrieve a document stored in the system, the requestor must enter a query for the document into the system. This may be done through a keyboard, for example. The queries used with the preferred embodiment of this system can be a natural language statement or string of words that describes the item. The search argument is created by testing the query words against the word index file. In many cases, the words in the search argument will occur in the key word records (abstracts) of several documents. In order to provide better discrimination between contending documents, different weights are applied to different key words. Weighting criteria are applied according to these general rules:

1 - Matches on numeric key words are given greater weight than matches on alpha key words.

2 - Matches with key words that are proper names or acronyms are given greater weight than matches with nouns or single purpose adjectives that are found in the dictionary memory.

3 - The weight assigned to a key word match is proportional to the number of times that the word occurred in the document divided by the log of the number of pages in the document.

4 - Matches with key words that occur in the first ten lines of the document are given greater weight than those of key words in the center of the body of text.

5 - Matches that occur with key words in the last five lines of text (before any copy lists) are given more weight than matches with words in the center of the text, but less weight than matches with words in the first ten lines.

6 - The weight of a key word match is increased when that word is the name of a month or year.

7 - The weight of a key word match is inversely proportional to the number of documents in the entire file that contain that key word in the body of the document (excluding occurrences as part of the copy list).

The rationale behind these general rules is to give the greatest weight to those matches that involve key words that have the most narrowly specific meaning. It is assumed that specific names, numbers and dates have very specific meaning so they are weighed heavily. It is also assumed that the most specific items will be mentioned at the beginning or end of the correspondence. Hence, words occurring in these regions are also given greater weight. An example of an expression that satisfies the general rules is the following:

Match Value =

$$\Sigma_{i,j} \frac{F_{i,j}+10^{A_i}+10^{K_i}+10^{L_i}+5^{E_i}+5^{H_i}}{\log_2 D_i} (1.25)^{M_i}(1.25)^{Y_i}$$

where:

$F_{i,j}$ = number of times ith key word appears in jth document divided $\log_2$ of the number of pages in document.

A_i = binary indicator if ith key word is an acronym or proper name.

K_i = binary indicator if ith key word occurs in first 10 lines.

L_i = binary indicator if ith key word is a numeric.

E_i = binary indicator if ith key word occurs in last 5 lines.

H_i = binary indicator if ith key word occurs in the dictionary as a noun or single purpose adjective.

M_i = binary indicator if ith key word is a month.

Y_i = binary indicator if ith key word is a year.

D_i = number of documents that contain ith key word.

Referring to FIG. 3, a flow chart of the processing of a query for a document is shown. At block 30 the user query is input to the processor 10 (FIG. 1) from input register 16 over bus 15.

The Query routine compares the query words to the contents of the word index file as shown in block 31 of the flow diagram of FIG. 3. The query words that match the word index file are processed at block 32 of the flow diagram by the Query Word Process routine.

Each query word is tested to determine if it is a month, year, numeric, acronym or normal (noun or single purpose adjective). A routine also adds weighting factors if the indicators in the word index fil3 show the word occurs in the first ten lines (Header) of the document, last five lines (Trailer) of the document, or occurs more than once in the document. Thus, a retrieval value of the word is calculated (block 33 of figure 3). The value of the word is reduced if it occurs in the copy list of the document or occurs in more than one document. An overall calculation of value for each word is calculated and a total value for all query words that match words in the word index file for each document number having any matches is accumulated. The steps of calculating the retrieval value for documents are shown in block 34 of FIG. 3. Following processing of all words in the query, the Query routine branches to the Month/Year Evaluation routine.

The Month/Year Evaluation routine increases the retrieval value for each document that contains a year and/or month that matches a year and/or month in the query. This routine then controls the processor 10 to output those documents from main memory 12 to output register 18 whose retrieval value is within 25 percent of the highest retrieval value calculated (block 35 of figure 3). Control is then returned to the Query routine which terminates the query procedure.

CLAIMS

1. Method for automatically abstracting and storing a document in machine readable form, characterized by the steps of :

   a) storing a dictionary of language terms commonly used in document preparation;

   b) appending codes to the language terms in said dictionary of language terms to identify selected parts of speech;

   c) comparing the language terms in an input document with the stored dictionary of language terms;

   d) selecting language terms from said input document which do not compare to the stored dictionary of language terms;

   e) selecting language terms from said input document which compare with language terms in said stored dictionary of language terms identified as selected parts of speech;

   f) coding the selected language terms with the identity of the input document; and

   g) storing the selected language terms for later recall.

2. Method according to Claim 1 further including the steps of accumulating a count for the number of times each of the selected language terms occurs in the input document and accumulating a count of the number of pages in the input document.

3. Method according to Claim 1 or Claim 2 further including the step of appending to each selected language term a code indicating the position of occurrence of the selected language term in the input document.

4. A method for retrieving a document from storage in response to input language terms descriptive of the content of the document characterized by the following steps :

a) comparing each of the input language terms to stored document abstract files of language terms, each document abstract language term having associated with it a code identifying its part of speech, a count indicating its frequency of occurrence in the document, a count of the number of pages in the document, and an indicator of the position of occurrence of the term in the document;

b) accumulating a retrieval record for each document abstract file composed of the language terms that compare equal;

c) calculating a document retrieval value for each retrieval record using the part of speech code, frequency count, number of pages in the document, and position indicator for each language term in the retrieval record;

d) increasing the document retrieval value for each retrieval record that includes a month and/or year; and

e) selecting the document corresponding to the highest calculated retrieval value for output.

5. Method according to Claim 4 further including the step of selecting all documents whose calculated retrieval value is equal to or greater than a predetermined percentage of the highest calculated retrieval value.

6. A system for automatically abstracting and storing a document in machine readable form comprising a memory, means for storing a dictionary of language terms commonly used in document preparation, said language terms including a code identifying certain ones of said language terms as selected parts of speech, means for receiving an input document of language terms in machine readable form, said input document including an identification code, and control means connected to said memory, said means for storing and said means for receiving,

said control means being characterized in that it comprises :

means for comparing the language terms of said input document to said dictionary of language terms,

means for selecting the language terms from said input document that compare unequal,

means for selecting the language term from said input document that compare equal and are coded as selected parts of speech;

means for counting the frequency of occurrence of the selected language terms in the input document;

means for counting the number of pages in the document;

means for calculating the position of occurrence of the selected language terms in the input document; and means for storing in said memory a record of each selected language term including the document identification code, the language term, the selected part of speech code, the frequency of occurrence count, the count of pages in the document, and the position of occurrence code.

7. System according to Claim 6 wherein said control means further includes means for comparing each selected language term from the input document to selected language terms currently stored in said memory, and means responsive to an equal compare for adding to the record of the selected language term stored in said memory the identification code of the input document, the frequency of occurrence count, and position of occurrence code for the selected language term, thereby eliminating the need for duplicate storage of the selected language.

8. System for retrieving a document from storage in response to an input query of language terms descriptive of the content of the document comprising a memory having stored therein language term records including the language term, identification codes of documents containing the language term, a selected parts of speech code, a frequency of occurrence count for the language term, a count of pages in each document, and a position of occurrence code for each document identification code in each language term record; said system being characterized in that it comprise :

means for comparing the language terms of the input query to language term records stored in said memory;

means for accumulating a retrieval record for each document identification code of each language term that compares equal;

means for calculating a document retrieval value for each retrieval record using the selected part of speech code, frequency of occurrence count, count of pages and position of occurrence code; and

means for outputting from memory the document whose identification code corresponds to the identification code for the highest calculated retrieval value.

9. System according to Claim 8 wherein said means for calculating further includes means for increasing the document retrieval value for each retrieval record that includes a month that compares equal to a term in the input query and further increasing the document retrieval value for each record that includes a year that compares equal to a term in the input query.

10. System according to Claim 8 or Claim 9 wherein said means for calculating includes means calculating a percentage of the highest calculated retrieval value for each document identification code, and said means for outputting further includes means for outputting all documents whose retrieval value exceeds a predetermined percentage of the highest calculated retrieval value.

11. System according to Claim 10 wherein means for outputting further includes means for selecting documents for display in the descending order of the number of query terms that matched language term records for the document.

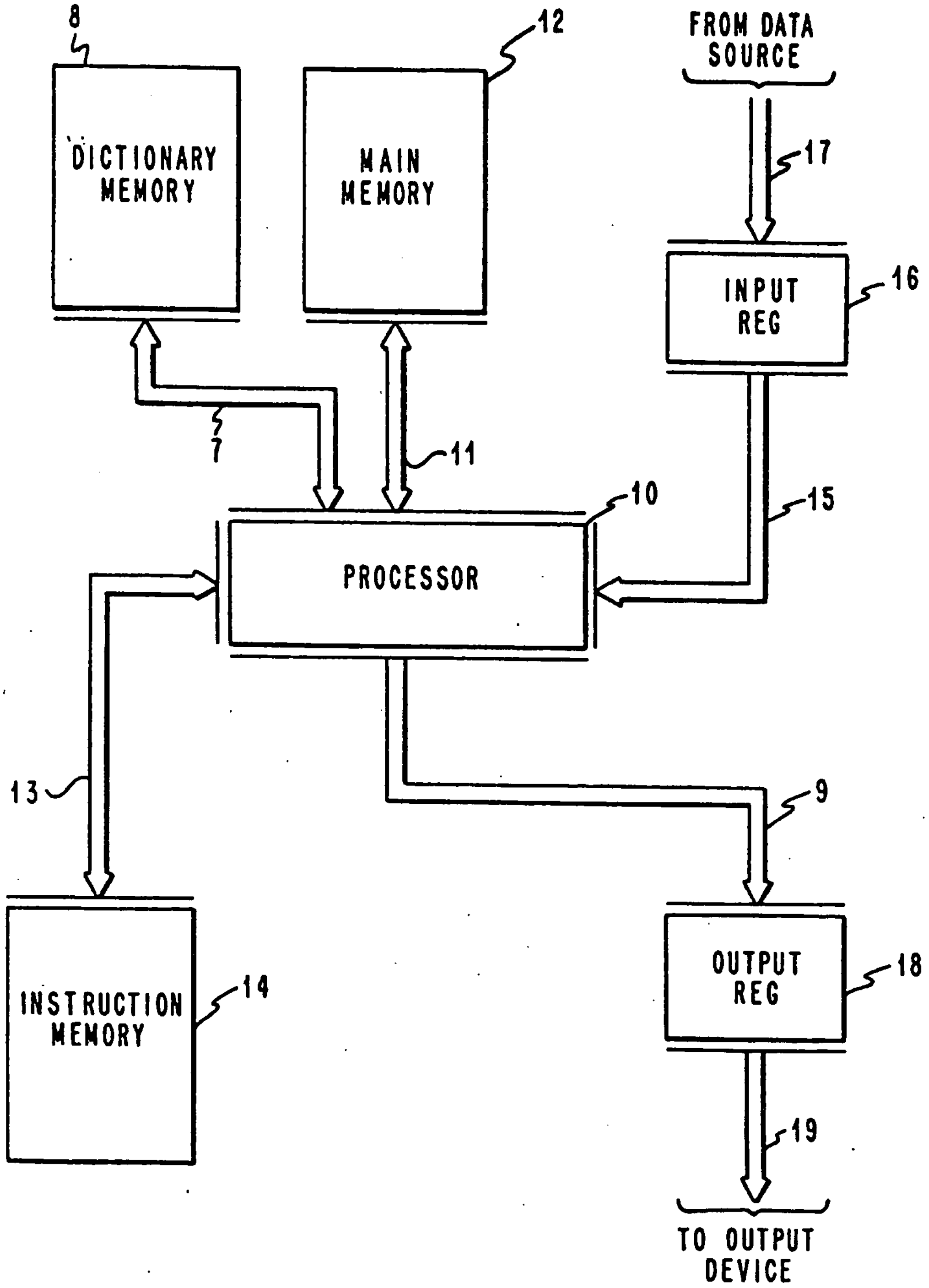
8
DICTIONARY MEMORY
12
MAIN MEMORY
FROM DATA SOURCE
17
16
INPUT REG
7
11
10
PROCESSOR
15
13
9
INSTRUCTION MEMORY
14
OUTPUT REG
18
19
TO OUTPUT DEVICE

FIG. 1

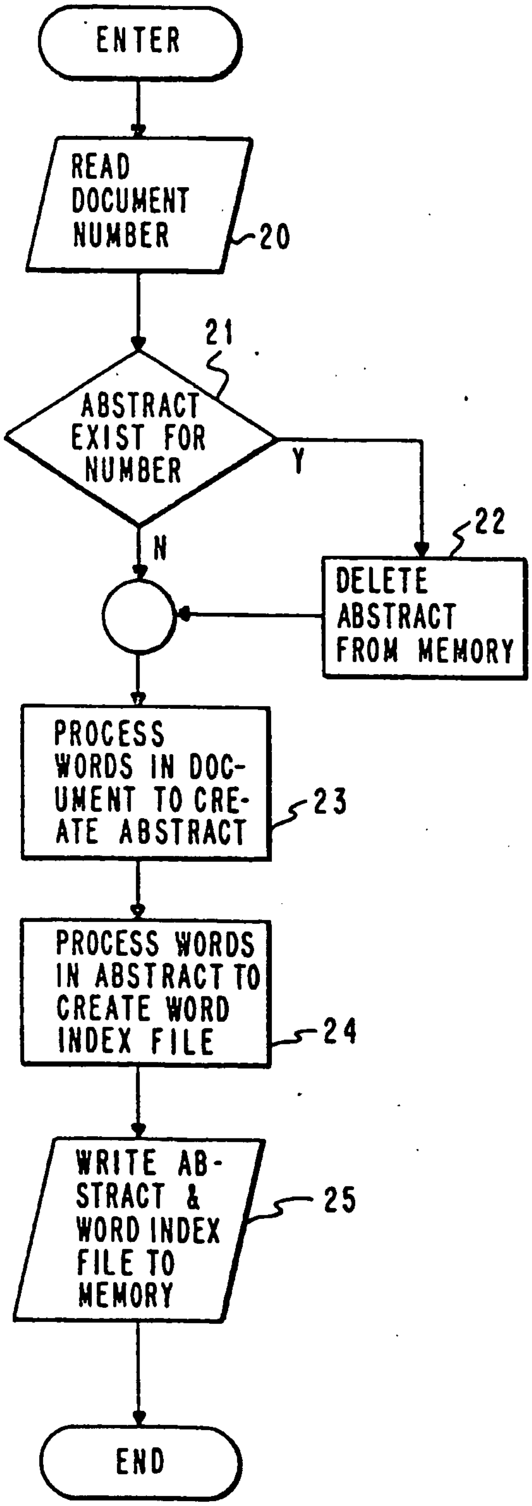
ENTER
READ DOCUMENT NUMBER
20
21
ABSTRACT EXIST FOR NUMBER
Y
N
22
DELETE ABSTRACT FROM MEMORY
PROCESS WORDS IN DOCUMENT TO CREATE ABSTRACT
23
PROCESS WORDS IN ABSTRACT TO CREATE WORD INDEX FILE
24
WRITE ABSTRACT & WORD INDEX FILE TO MEMORY
25
END

FIG. 2

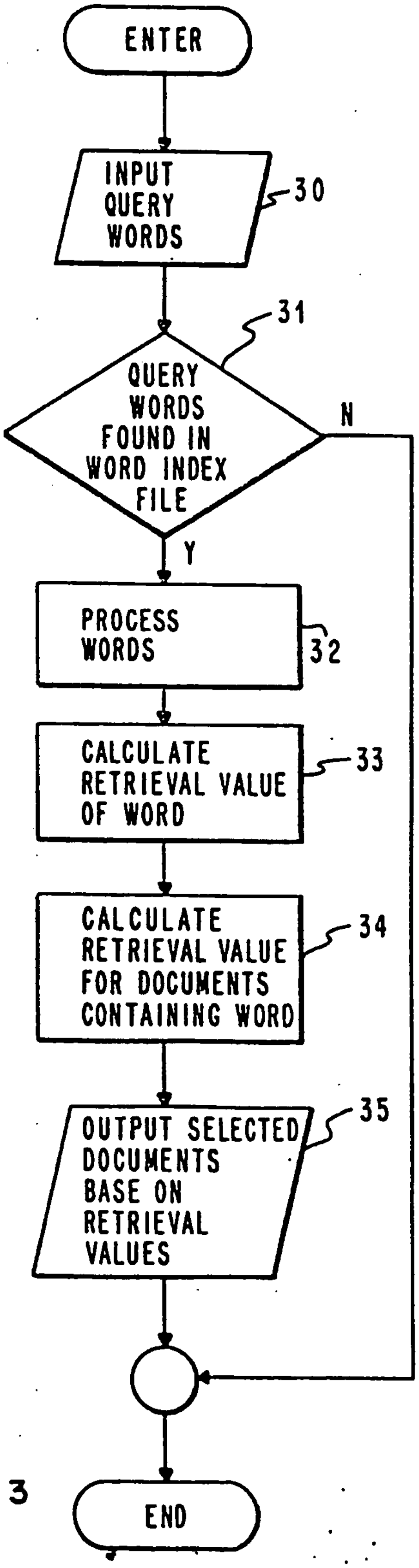
ENTER
INPUT QUERY WORDS
30
31
QUERY WORDS FOUND IN WORD INDEX FILE
N
Y
PROCESS WORDS
32
CALCULATE RETRIEVAL VALUE OF WORD
33
CALCULATE RETRIEVAL VALUE FOR DOCUMENTS CONTAINING WORD
34
OUTPUT SELECTED DOCUMENTS BASE ON RETRIEVAL VALUES
35
END

FIG. 3

0032194

European Patent Office

EUROPEAN SEARCH REPORT

Application number
EP 80 10 7625

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim
X	PROCEEDINGS OF THE SPRING JOINT COMPUTER CONFERENCE (AFIPS CONFERENCE PROCEEDINGS), vol. 34, May 14-16, 1969, MONTRALE (US) HILLMAN et al.: "The LEADER Retrieval System", pages 447-455 * page 451, left-hand column, line 7 to page 455, left-hand column, line 35 * ---	1,3,4,6,7
	JOURNAL OF THE ASSOCIATION FOR COMPUTING MACHINERY, vol. 20, no. 2, April 1973, NEW YORK (US) SALTON: "Recent Studies in Automatic Text Analysis and Document Retrieval", pages 258-278 * Page 261, lines 14-24; page 266, line 9 to page 267, line 44; Tables III-V * ---------	1,2,4-8,10,11

CLASSIFICATION OF THE APPLICATION (Int. Cl.³)

G 06 F 15/40

TECHNICAL FIELDS SEARCHED (Int. Cl.³)

G 06 F 15/20
15/38
15/40
G 06 K 9/72

CATEGORY OF CITED DOCUMENTS

X: particularly relevant
A: technological background
O: non-written disclosure
P: intermediate document
T: theory or principle underlying the invention
E: conflicting application
D: document cited in the application
L: citation for other reasons

&: member of the same patent family, corresponding document

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	28.03.1981	HARRIS